

INCREASED SENSITIVITY IN FMRI GROUP ANALYSIS USING MIXED-EFFECT MODELING

Merlin Keller¹, Alexis Roche²

¹INRIA / ²CEA, Neurospin, Gif-sur-Yvette, France
Institut d'Imagerie Neurofonctionnelle (IFR 49), Paris, France

ABSTRACT

In functional Magnetic Resonance Imaging group studies, uncertainties on the individual BOLD responses are not taken into account by standard detection procedures, which may limit their sensitivity. Mixed-effect models have been introduced to derive decision statistics that weight the subjects according to their reliability. To date, however, the associated statistical tests are almost not used by investigators, partly because they are inexact in that they control only approximately the false positive risk. We tackle this problem using a permutation testing framework that yields exact tests under mild nonparametric assumptions. This approach enables us to evaluate the sensitivity of mixed-effect statistics on a mental calculation experiment involving men and women.

Index Terms— Magnetic resonance, Brain, Statistics.

1. INTRODUCTION

The object of functional Magnetic Resonance Imaging (fMRI) group analysis is to identify brain regions where the Blood Oxygenation Level Dependent (BOLD) responses correlate with behavioral, clinical, or genetic predictors. For instance, one may seek differences in BOLD responses between mentally ill patients and healthy controls, or correlate responses with psychological test scores in an homogeneous population. Group inferences are usually performed via massively univariate t - or F -tests [1] after a spatial normalization step that renders data from the different subjects grossly comparable on a voxelwise basis. The general linear model (GLM) underlying both t - and F -tests assumes that the normalized data is identically and normally distributed across subjects.

This condition, however, is barely met due to different amounts of noise in the fMRI time series, subject-dependent deviations from the canonical hemodynamic response function, spatial normalization errors, etc. We may thus expect increased sensitivity by extending the GLM so as to account for inhomogeneous errors (heteroscedasticity), and derive test statistics accordingly. In this paper, we generalize the maximum likelihood ratio-based approach of [2], which was restricted to one-sample inference. When testing the global dependence between the BOLD responses and the predictors,

the maximum likelihood ratio statistic may be calibrated using permutations to yield an exact test under mild assumptions, unlike previous parametric approaches [3, 4, 5, 6].

2. MIXED-EFFECT MODEL

After scanning n subjects during a cognitive experiment, we process respective fMRI data individually so that, in each particular voxel of the reference grid and for each subject i , we have a noisy estimate y_i of the BOLD effect in response to a given contrast of experimental conditions. Provided that a large number of scans is available [3, 2], we shall assume that y_i is normally distributed around the unobserved effect z_i , that is: $y_i = z_i + e_i$ with $e_i \sim \mathcal{N}(0, s_i^2)$, the standard error s_i being known.

Our goal is to correlate the effects $\mathbf{z} = [z_1, \dots, z_n]^\top$ with a given set of p predictors, as represented by a $n \times p$ matrix \mathbf{X} . We start with assuming that \mathbf{z} relates with \mathbf{X} through a GLM: $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and the error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{V})$ is a zero-mean multivariate Gaussian. Further assuming statistical independence between the within-subject and between-subject variability sources, we get:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}', \quad \boldsymbol{\varepsilon}' \sim \mathcal{N}(0, \mathbf{V} + \mathbf{W}),$$

with $\mathbf{W} = \text{diag}(s_1^2, \dots, s_n^2)$. In our setting, \mathbf{W} is known while \mathbf{V} is searched in the space \mathcal{S} of positive scalar matrices, *i.e.* $\mathbf{V} = \sigma^2 \mathbf{I}_n$ for some $\sigma > 0$, meaning that the effects are assumed independently and identically distributed. This model is both a special case of Laird & Ware's two-stage linear model [7], and a generalization of the standard GLM, which corresponds to $\mathbf{W} = 0$.

2.1. Parameter estimation

In order to estimate both the effect $\boldsymbol{\beta}$ and the variance \mathbf{V} , we may jointly minimize the negated log-likelihood:

$$\begin{aligned} -2 \log p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{V}) &= n \log(2\pi) + \log |\mathbf{V} + \mathbf{W}| \\ &+ \|(\mathbf{V} + \mathbf{W})^{-\frac{1}{2}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 \end{aligned}$$

At fixed \mathbf{V} , this yields a weighted least-square problem. However, since \mathbf{V} is also to be optimized, there is no general closed-form solution. In order to find at least a local likelihood maximizer, we use the Expectation-Maximization (EM) algorithm [8] that derives from considering \mathbf{z} as missing data. At each iteration k , the following two steps are performed:

◦ *E-step*. Given current estimates (β_k, \mathbf{V}_k) , compute the posterior density $p(\mathbf{z}|\mathbf{y}, \beta_k, \mathbf{V}_k)$, which is seen to be normal $\mathcal{N}(\bar{\mathbf{z}}_k, \Sigma_k)$ with:

$$\begin{aligned}\Sigma_k &= (\mathbf{W}^{-1} + \mathbf{V}_k^{-1})^{-1} \\ \bar{\mathbf{z}}_k &= \Sigma_k(\mathbf{W}^{-1}\mathbf{y} + \mathbf{V}_k^{-1}\mathbf{X}\beta_k)\end{aligned}$$

◦ *M-step*. Update the parameter estimates by maximizing the complete-data expected log-likelihood:

$$\mathcal{Q}_k(\beta, \mathbf{V}) \propto \log |\mathbf{V}| + \|\mathbf{V}^{-\frac{1}{2}}(\mathbf{z}_k - \mathbf{X}\beta)\|^2 + \text{trace}(\Sigma_k \mathbf{V}^{-1}),$$

yielding the explicit rule:

$$\begin{aligned}\beta_{k+1} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{z}}_k \\ \mathbf{V}_{k+1} &= \frac{1}{n} \left[\|\bar{\mathbf{z}}_k - \mathbf{X}\beta_{k+1}\|^2 + \text{trace}(\Sigma_k) \right] \mathbf{I}_n\end{aligned}$$

Notice that joint likelihood maximization is one of the two estimation methods originally proposed for the two-stage linear model [7]. Alternatively, \mathbf{V} may be estimated prior to β by restricted maximum likelihood (ReML), which is shown to yield an unbiased estimate and seems therefore preferable from a pure estimation perspective. There is no evidence, however, that the ReML strategy leads to more powerful testing procedures.

2.2. Generalized F statistic

We now turn to the problem of testing the null hypothesis $H_0: \mathbf{C}\beta = 0$ for a given $q \times p$ matrix \mathbf{C} . In standard GLM context, such a test classically uses a F statistic, which lacks justification under our more general model and may hence be sub-optimally sensitive. A systematic and customary way to select a test statistic is then to use the (log) maximum likelihood ratio (MLR):

$$\Lambda = -2 \log \frac{\max_{(\mathbb{R}^p \cap H_0) \times \mathcal{S}} p(\mathbf{y}|\beta, \mathbf{V})}{\max_{\mathbb{R}^p \times \mathcal{S}} p(\mathbf{y}|\beta, \mathbf{V})}$$

We may compute the denominator using the above EM algorithm. Computing the numerator involves a straightforward EM variant in which the M -step is modified so that β is optimized under the linear constraint $\mathbf{C}\beta = 0$, yielding the update rule: $\beta_{k+1} = \mathbf{P}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\mathbf{z}}_k$ where $\mathbf{P} = \mathbf{I}_p - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{C}^\top [\mathbf{C}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{C}^\top]^{-1} \mathbf{C}$.

Owing to a general result known as Wilks' phenomenon, Λ is asymptotically distributed like a χ_q^2 as the sample size n increases, which provides a quick approximate test. Interestingly, Λ generalizes the F statistic in the sense that, for a

standard GLM ($\mathbf{V} = \sigma^2 \mathbf{I}_n$ and $\mathbf{W} = 0$), both statistics are related through a strictly increasing function:

$$\Lambda = n \log \left[1 + \frac{qF}{n-p} \right] \underset{n \rightarrow \infty}{\sim} qF,$$

and are thus equivalent from a decision theoretic viewpoint.

2.3. Generalized t statistic

When $q = 1$ so that the contrast $\mathbf{C}\beta$ is a real number, one may want to perform a one-sided test in order to filter out negative contrasts. To that end, we define the following one-sided MLR variant: $\lambda = \text{sign}(\mathbf{C}\hat{\beta})\sqrt{\Lambda}$, where $\hat{\beta}$ is the maximum likelihood effect estimate as approximated by the EM algorithm. By Wilks' phenomenon, λ is asymptotically distributed like $\mathcal{N}(0, 1)$ under H_0 . Furthermore, it is easy to check that λ generalizes the t statistic given that $t = \text{sign}(\mathbf{C}\hat{\beta}_{ols})\sqrt{F}$ where $\hat{\beta}_{ols}$ is the ordinary least-square estimate.

2.4. Permutation test

Consider the special case where $\mathbf{C} = [\mathbf{I}_{p-1}, 0_{p-1 \times 1}]$, assuming conventionally that the last column of \mathbf{X} is the constant predictor. Testing $H_0: \mathbf{C}\beta = 0$ then amounts to testing the global statistical independence between \mathbf{z} and \mathbf{X} . Such a test can be performed approximately by calibrating the MLR statistic using the χ_{p-1}^2 law, however this may lead to biased false positive control for small samples. Permutation tests then offer a valuable alternative, being exact at all sample sizes and robust against deviations from normality [9, 10, 2].

The permutation test consists in tabulating the distribution of Λ (or virtually any test statistic) by shuffling the data according to $y_i \rightarrow y_{\pi(i)}$, for each permutation π of $\{1, \dots, n\}$, and computing the corresponding value Λ_π of the test statistic. This is justified by the fact that, under statistical independence, the observations are *exchangeable* in that all the $n!$ permuted samples are equally likely.

3. RESULTS

In the sequel, we focus on the two-sample model, in which \mathbf{X} has two columns: \mathbf{x}_1 is a binary vector of zeros and ones standing for group labels (e.g. males/females), and $\mathbf{x}_2 \equiv \mathbf{1}$ is the constant predictor used to model a group-independent baseline. To test the effect of group membership, we may then use the generalized t statistic associated with the one-dimensional contrast $\mathbf{C} = [1, 0]$, and calibrate the test using the above permutation mechanism which, in this case, simplifies to permutations of labels [9].

3.1. Simulation

We simulated data according to the model with a nonzero group effect $\beta_1 = 1$ and an offset $\beta_2 = 0$ (which has no

impact). The within-subject variances s_i^2 were generated independently from a $\Gamma(a, b)$ distribution, and the between-subject variance was set to $\sigma^2 = 1$.

In order to examine the effects of heteroscedasticity and sample size, we considered groups of equal size $N = 5, 10, 15, 20$, and Gamma distribution parameters (a, b) such that: $b = 1/4, 1, 4, a = 1/b$. This way, within-subject variances were distributed with mean 1 and variance b controlling the “degree of heteroscedasticity”. For all $4 \times 3 = 12$ possible combinations of N and b , 100 000 independent samples of size N were generated and the permutation tests associated with both the t statistic and our mixed-effect (MFX) generalization were performed.

By varying the detection threshold over the range of observed statistic values, receiving operator characteristic (ROC) curves were obtained in each case. Each ROC curve was then summarized by its area under curve (AUC), an overall measure of sensitivity. Figure 1 contains plots of AUC values against group sizes for every value of b and each test statistic.

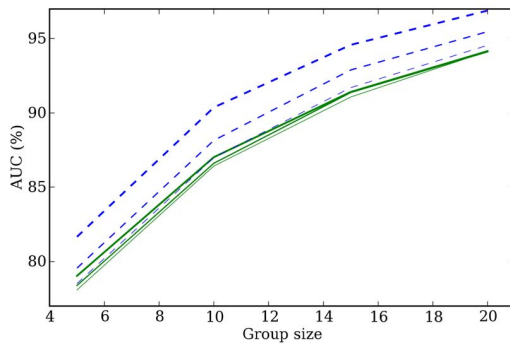


Fig. 1. Area under curve (AUC) of ROC curves obtained for different values of n and b , using the standard t statistic (solid line) and its MFX generalization (dashed line). Line thickness indicates the value of b : thin lines correspond to $b = 1/16$, medium-thick lines to $b = 1$, and thick lines to $b = 16$.

As would be expected, sensitivity as measured by the AUC is seen to increase with sample size. Theoretically, this quantity goes to 1 as the number of subjects becomes arbitrarily large, for both test statistics and for any value of b . However, even for datasets of 20 subjects, which is considered ample in fMRI group analysis context, a noticeable sensitivity gap can be observed between standard GLM-based and MFX tests, even under moderate heteroscedasticity.

3.2. Real data

We now present results of the method on a “localizer” fMRI dataset involving 10 subjects of each sex [11]. Among other tasks, the subjects had to perform mental calculations specified by oral instructions. Subtracting the effect of passively

hearing sentences from the global effect induced by the calculation task thus enables us to detect which brain regions are involved specifically in mental calculus. The question is whether this task is performed differently by men and women.

Within-subject analyses were conducted using SPM2 (Statistical Parametric Mapping software). Data were submitted successively to motion correction, slice timing, spatial normalization, and spatial smoothing using a $5 \times 5 \times 5\text{mm}^3$ FWHM Gaussian filter. At the between-subject level, we are interested in regions that display higher activation levels in one group than in the other, thus defining the “Female – Male” and “Male – Female” contrasts. Group analyses were performed on the intersection of the whole-brain masks of all subjects (36,806 voxels). For both contrasts, the statistical maps were thresholded for a 1% false positive rate using permutations.

For comparison, we also report the results of the parametric t -test implemented in SPM, when thresholded at the same level of expected false positive rate ($P \leq 0.01$ uncorrected), using the Student distribution with 18 degrees of freedom. In SPM, both voxel-level and cluster-level corrected P -values are computed using closed-form approximations based on Random Field Theory [12]. In the following, we report clusters whose cluster-level or voxel-level P -value was found less than 5% in at least one of the three statistical procedures.

“Female – Male” contrast. Table 1 and Fig. 2 summarize the results obtained for the “Female – Male” between-subject contrast. A single significant region is detected in the left intra-parietal sulcus by all three tests. The increasing values for cluster extent suggest a higher sensitivity of the permutation approach over the parametric one, and of the MFX t statistic over the usual t statistic. The latter observation is confirmed by the corrected P -values, which are more significant for the MFX test, especially at the voxel level in this case.

The region is found significant at 5% in terms of cluster size by the SPM t -test, and in terms of activation peak by the MFX permutation test, while it falls short of significance according to both criteria with the permutation test using the usual t statistic. It must be stressed however that the significance levels given by SPM are known to be biased when the applicability conditions of random field theory are not met [13], which is the case here given the strong disagreement with permutation-based P -values which are theoretically exact.

“Male – Female” contrast. Lower halves of Table 1 and Fig. 2 summarize the results obtained for the “Male – Female” contrast. A single significant region is detected in the left angular gyrus, this time only by the permutation MFX test. While not significant at the cluster level, its activation peak survives voxel-level familywise error correction at 5%. It can also be noted that the cluster-level P -value given by

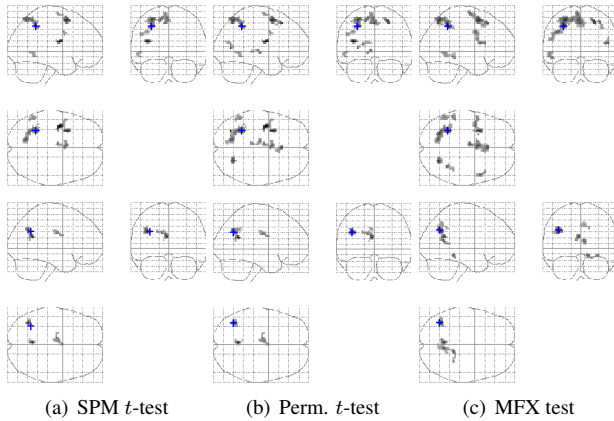


Fig. 2. Maximum Intensity Projection (MIP) of the group statistical maps obtained for “Female – Male” (top row) and “Male – Female” (bottom row) in a mental calculus task. Only clusters larger than 20 voxels are represented.

Cluster anatomical location	Statistical test procedure	Cluster-level P_{corr}	Cluster extent (voxels)	Voxel peak P_{corr}
Female – Male				
Left Intra-Parietal Sulcus	SPM t -test	0.01	99	1
	Perm. t -test	0.13	103	0.68
	MFX test	0.09	155	0.04
Male – Female				
Left Angular Gyrus	SPM t -test	0.67	30	0.99
	Perm. t -test	0.50	32	0.51
	MFX test	0.42	43	0.04

Table 1. Results of two-sample inference for a mental calculus task.

SPM is now more conservative than that given by the permutation tests, presumably because of the small size of the detected cluster.

4. CONCLUSION

Our study demonstrates the possibility of increasing sensitivity in group analyses while maintaining exact control on specificity. This is done by combining a mixed-effect variant of the t (or F) statistic with a permutation test. From a cognitive viewpoint, the proposed test was the only one we tried to reveal that, during a mental calculus task, women might have higher tendency than men to solicit the left intra-parietal sulcus, known to be involved in number processing. On the other hand, men might have a higher level of activation in the left angular gyrus, known to be involved in memory recalling.

5. REFERENCES

- [1] K. J. Friston, *Human Brain Function*, chapter 2, pp. 25–42, Academic Press, 1997.
- [2] S. Mériaux, A. Roche, G. Dehaene-Lambertz, B. Thirion, and J.-B. Poline, “Combined permutation test and mixed-effect model for group average analysis in fMRI,” *Hum. Brain Mapp.*, vol. 27, no. 5, pp. 402–410, 2006.
- [3] K.J. Worsley, C.H. Liao, J. Aston, V. Petre, G.H. Duncan, F. Morales, and A.C. Evans, “A general statistical analysis for fMRI data,” *Neuroimage*, vol. 15, no. 1, pp. 1–15, 2002.
- [4] K.J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, “Classical and bayesian inference in neuroimaging: Theory,” *Neuroimage*, vol. 16, no. 2, pp. 465–483, 2002.
- [5] C.F. Beckmann, M. Jenkinson, and S.M. Smith, “General multi-level linear modelling for group analysis in fMRI,” *Neuroimage*, vol. 20, pp. 1052–1063, 2003.
- [6] M. Woolrich, T. Behrens, C. Beckmann, M. Jenkinson, and S. Smith, “Multi-level linear modelling for fMRI group analysis using Bayesian inference,” *Neuroimage*, vol. 21, no. 4, pp. 1732–1747, 2004.
- [7] N. M. Laird and J. H. Ware, “Random-Effects Models for Longitudinal Data,” *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.
- [8] A.P. Dempster, A.P. Laird, and D.B. Rubin, “Maximum likelihood from incomplete data via the em algorithm (with discussion),” *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.
- [9] Phillip Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer, 3rd edition, 2005.
- [10] T.E. Nichols and A.P. Holmes, “Nonparametric permutation tests for functional neuroimaging: A primer with examples,” *Hum. Brain Mapp.*, vol. 15, pp. 1–25, 2002.
- [11] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J.-B. Poline, “Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses,” *Neuroimage*, vol. 35, no. 1, pp. 105–120, 2007.
- [12] K.J. Worsley, “Local maxima and the expected Euler characteristic of excursion sets of χ^2 , f , and t fields,” *Adv. Appl. Prob.*, vol. 26, pp. 13–42, 1994.
- [13] S. Hayasaka and T.E. Nichols, “Validating Cluster Size Inference: Random Field and Permutation Methods,” *Neuroimage*, vol. 20, no. 4, pp. 2343–2356, 2003.